Retrieval-Augmented Generation for AI-Generated Content: A Survey

Abstract

Advancements in model algorithms, the growth of foundational models, and access to high-quality datasets have propelled the evolution of Artificial Intelligence Generated Content (AIGC). Despite its notable successes, AIGC still faces hurdles such as updating knowledge, handling long-tail data, mitigating data leakage, and managing high training and inference costs. Retrieval-Augmented Generation (RAG) has recently emerged as a paradigm to address such challenges. In particular, RAG introduces the information retrieval process, which enhances the generation process by retrieving relevant objects from available data stores, leading to higher accuracy and better robustness. In this paper, we comprehensively review existing efforts that integrate RAG technique into AIGC scenarios. We first classify RAG foundations according to how the retriever augments the generator, distilling the fundamental abstractions of the augmentation methodologies for various retrievers and generators. This unified perspective encompasses all RAG scenarios, illuminating advancements and pivotal technologies that help with potential future progress. We also summarize additional enhancements methods for RAG, facilitating effective engineering and implementation of RAG systems. Then from another view, we survey on practical applications of RAG across different modalities and tasks, offering valuable references for researchers and practitioners. Furthermore, we introduce the benchmarks for RAG, discuss the limitations of current RAG systems, and suggest potential directions for future research. Github: https://github.com/PKU-DAIR/RAG-Survey.